Beyond Sentence-Level End-to-End Speech Translation: Context Helps

Beyond Sentence-Level End-to-End Speech Translation: Context Helps

Biao Zhang[1] **Ivan Titov**[1,2] **Barry Haddow**[1] **Rico Sennrich**[3,1]

[1]School of Informatics, University of Edinburgh
[2]ILLC, University of Amsterdam
[3]Department of Computational Linguistics, University of Zurich
B.Zhang@ed.ac.uk, {ititov,bhaddow}@inf.ed.ac.uk, sennrich@cl.uzh.ch

Abstract

Document-level contextual information has shown benefits to text-based machine translation, but whether and how context helps end-to-end (E2E) speech translation (ST) is still under-studied. We fill this gap through extensive experiments using a simple concatenation-based context-aware ST model, paired with adaptive feature selection on speech encodings for computational efficiency. We investigate several decoding approaches, and introduce in-model ensemble decoding which jointly performs document- and sentence-level translation using the same model. Our results on the MuST-C benchmark with Transformer demonstrate the effectiveness of context to E2E ST. Compared to sentence-level ST, context-aware ST obtains better translation quality (+0.18-2.61 BLEU), improves pronoun and homophone translation, shows better robustness to (artificial) audio segmentation errors, and reduces latency and flicker to deliver higher quality for simultaneous translation.[1]

1 Introduction

Document-level context often offers extra informative clues that could improve the understanding of individual sentences. Such clues have been proven effective for textual machine translation (MT), particularly in handling translation errors specific to discourse phenomena, such as inaccurate coreference of pronouns (Guillou, 2016) and mistranslation of ambiguous words (Rios et al., 2017). Besides, ensuring consistency in translation is virtually impossible without document-level context as well (Voita et al., 2019). Analogous to MT, speech translation (ST) also suffers from these translation issues, and super-sentential context could in fact be more valuable to ST because 1) homophones and acoustic noise bring additional ambiguity to ST, and 2) a common use case in ST is simultaneous translation, where the system has to output translations of sentence fragments, and may have to predict future input to account for word order differences between the source and target language (Grissom II et al., 2014). Both for ambiguity from the acoustic signal, and operating on small sentence fragments, we hypothesize that access to extra context[2] will be beneficial.

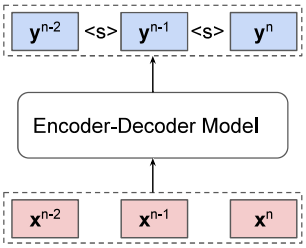

Figure 1: Overview of the concatenation-based context-aware ST. $\mathbf{y}^n$ denotes the n-th target sentence in a document; $\mathbf{x}^n$ denotes the speech encodings extracted from the n-th audio segment. We use dashed gray box to indicate the concatenation operation. "<s>": sentence separator symbol.

Although recent studies on ST have achieved promising results with end-to-end (E2E) models (Anastasopoulos and Chiang, 2018; Di Gangi et al., 2019; Zhang et al., 2020a; Wang et al., 2020; Dong et al., 2020), nevertheless, they mainly focus on sentence-level translation. One practical challenge when scaling up sentence-level E2E ST to the document-level is the encoding of very long audio segments, which can easily hit the computational bottleneck, especially with Transformers (Vaswani et al., 2017). So far, the research question of whether and how contextual information benefits E2E ST has received little attention.

In this paper, we answer this question through extensive experiments by exploring a concatenation-

[1]Source code is available at https://github.com/bzhangGo/zero.

[2]By default, we use *context* to denote both source- and target-side information from previous sentences.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2566–2578
August 1–6, 2021.

based context-aware ST model. Figure 1 illustrates our model, where neighboring source (target) sequences are chained together into one sequence for joint translation. This paradigm only requires data-level manipulation, thus allowing us to reuse any existing sentence-level E2E ST models. Despite its simplicity, this approach successfully leverages contextual information to improve textual MT (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Lopes et al., 2020), and here we adapt it to ST. As for the computational bottleneck, we shorten the speech encoding sequence via adaptive feature selection (Zhang et al., 2020b,a, AFS), which only retains a small subset of encodings (∼16%) for each audio segment.

We investigate several decoding methods, including chunk-based decoding and sliding-window based decoding. We also study an extension of the latter with the constraint of target prefix, where the prefix denotes the translation of previous context speeches. We find that using these methods sometimes results in *misaligned translations*, particularly when using the constraint. This issue manifests itself in mismatching sentence boundaries and producing over- and/or under-translation, which greatly hurts sentence-based evaluation metrics. To avoid such misalignments, we introduce *in-model ensemble decoding* (IMED) to regularize the document-level translation with its sentence-level counterpart. Note that we use the same context-aware ST model here for both types of translation – that's why we call it *in-model* ensemble.

We adopt Transformer (Vaswani et al., 2017) for experiments with the MuST-C dataset (Di Gangi et al., 2019). We study the impact of context on translation in different settings. Our results demonstrate the effectiveness of contextual modeling. Our main findings are summarized below:

- Incorporating context improves overall translation quality (+0.18-2.61 BLEU) and benefits pronoun translation across different language pairs, resonating with previous findings in textual MT (Miculicich et al., 2018; Huo et al., 2020). In addition, context also improves the translation of homophones.
- ST models with contexts suffer less from (artificial) audio segmentation errors.
- Contextual modeling improves translation quality and reduces latency and flicker for simultaneous translation under re-translation strategy (Arivazhagan et al., 2020a).

2 Related Work

Our work is inspired by pioneer studies on context-aware textual MT. Context beyond the current sentence carries information whose importance for translation cohesion and coherence has long been posited (Hardmeier et al., 2012; Xiong and Zhang, 2013). With the rapid development of neural MT and also available document-level textual datasets, research in this direction gained great popularity. Recent efforts often focus on either advanced contextual neural architecture development (Tiedemann and Scherrer, 2017; Kuang et al., 2018; Miculicich et al., 2018; Zhang et al., 2018, 2020c; Kang et al., 2020; Chen et al., 2020; Ma et al., 2020a; Zheng et al., 2020) and/or improved analysis and evaluation targeted at specific discourse phenomena (Bawden et al., 2018; Läubli et al., 2018; Guillou et al., 2018; Voita et al., 2019; Kim et al., 2019; Cai and Xiong, 2020). We follow this research line, and adapt the concatenation-based contextual model (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Lopes et al., 2020) to ST. Our main interest lies in exploring the impact of context on ST. Developing dedicated contextual models for ST is beyond the scope of this study, which we leave to future work.

Context-aware ST extends the sentence-level ST towards streaming ST which allows models to access unlimited previous audio inputs. Instead of improving contextual modeling, many studies on streaming ST aim at developing better sentence/-word segmentation policies to avoid segmentation errors that greatly hurt translation (Matusov et al., 2007; Rangarajan Sridhar et al., 2013; Iranzo-Sánchez et al., 2020; Zhang and Zhang, 2020; Arivazhagan et al., 2020b). Very recently, Ma et al. (2020b) proposed a memory augmented Transformer encoder for streaming ST, where the previous audio features are summarized into a growing continuous memory to improve the model's context awareness. Despite its success, this method ignores the target-side context, which turns out to have significant positive impact on ST in our experiments.

Our study still relies on *oracle* sentence segmentation of the audio. The most related work to ours is (Gaido et al., 2020), which also investigated contextualized translation and showed that context-aware ST is less sensitive to audio segmentation errors. While they exclusively focus on the robustness to segmentation errors, our study investigates the benefits of context-aware E2E ST more broadly.

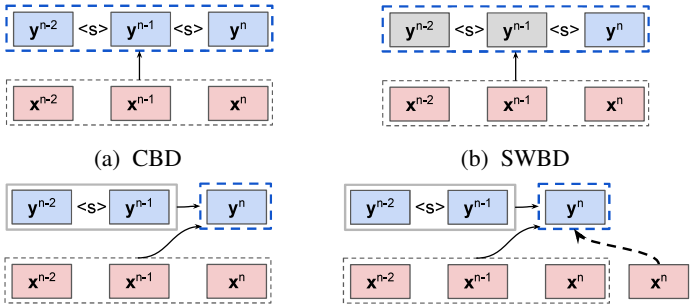

Figure 2: Illustration of different decoding methods: chunk-based decoding (CBD, 2a), sliding-window based decoding without (SWBD, 2b) and with (SWBD-Cons, 2c) the target prefix constraint and the proposed in-model ensemble decoding (IMED, 2d). The dashed blue box denotes model generation; the solid gray box (2c, 2d) indicates the target prefix constraint; sentences in the gray rectangle (2b) are discarded after generation. The dashed arrow in IMED stands for the sentence-level translation.

3 Context-aware ST via Concatenation

We extend the sentence-level ST with document-level context, by modeling up to C previous source/target segments/sentences for translation. Formally, given a pre-segmented audio (source document) $\mathbf{A} = \left(\mathbf{a}^1, \ldots, \mathbf{a}^N\right)$ as well as its paired target document $\mathbf{Y} = \left(\mathbf{y}^1, \ldots, \mathbf{y}^N\right)$, the model is trained to maximize the following likelihood:

$$\log p\left(\mathbf{Y}|\mathbf{A}\right) = \sum_{n=1}^{N} \log p\left(\mathbf{y}^n|\mathbf{x}^n, \mathcal{C}_{\mathbf{y}}^n, \mathcal{C}_{\mathbf{x}}^n\right), \quad (1)$$

where $\mathbf{x}^n = \text{AFS}\left(\mathbf{a}^n\right)$, i.e. the speech encodings extracted via AFS (Zhang et al., 2020a). $\mathbf{a}^n$ and $\mathbf{y}^n$ denote the n-th audio segment and target sentence, respectively. N is the number of segments/sentences in the document. $\mathcal{C}_{\mathbf{x}}^n$ and $\mathcal{C}_{\mathbf{y}}^n$ stand for the source and target context, respectively, i.e. $\{\mathbf{x}^{n-i}\}_{i=1}^{C}$ and $\{\mathbf{y}^{n-i}\}_{i=1}^{C}$.

Adaptive Feature Selection Audio segment is often converted into frame-based features for neural modeling. Different from text, each segment might contain hundreds or even thousands of such features, making contextual modeling computationally difficult. Zhang et al. (2020a) found that most speech encodings emitted by a Transformer-based audio encoder carry little information for translation, and their deletion even improves translation quality. We follow Zhang et al. (2020a) and perform AFS to only extract those informative encodings ($\sim$16%) optimized via sentence-level speech recognition with $\mathcal{L}_0$DROP (Zhang et al., 2020b). This greatly shortens the speech encoding sequence, thus enabling broader context exploration.

Concatenation-based Contextual Modeling We adopt the concatenation method to incorporate the previous context ($\mathcal{C}_{\mathbf{x}}^n$/$\mathcal{C}_{\mathbf{y}}^n$) (Tiedemann and Scherrer, 2017; Bawden et al., 2018) as shown in Figure 1. After obtaining the AFS-based encodings ($\mathbf{x}^n$) for each audio segment, we concatenate those encodings of neighboring segments to form the source input. The same is applied to the target-side sentences, except for a separator symbol "<s>" inserted in-between sentences to distinguish sentence boundaries.[3] Such modeling enables us to use arbitrary encoder-decoder models for context-aware ST, such as the Transformer (Vaswani et al., 2017) used in this paper. Despite no dedicated hierarchical modeling (Miculicich et al., 2018), this paradigm still allows for intra- and inter-sentence attention during encoding and decoding, which explicitly utilizes context for translation and has been proven successful (Lopes et al., 2020).

4 Inference

Concatenation-based contextual modeling allows for different inference strategies with possible trade-offs between simplicity/efficiency and accuracy. We investigate the following inference strategies (see Figure 2):

Chunk-based Decoding (CBD) CBD splits all audio segments in one document into non-overlapping chunks, with each chunk concatenating $C + 1$ segments, as shown in Figure 2a. CBD directly translates each chunk, and then recovers sentence-level translation via the separator symbol "<s>". CBD is the most efficient inference strategy, only encoding/decoding each sentence once, but it might suffer from *misaligned translation*,

[3] Note that we did not add similar boundary information to audio segments, because AFS implicitly captures these signals through independent segment encoding.

producing more or fewer sentences than the input segments. We simply drop the extra generated sentences and replace the missing ones with "<unk>" when computing sentence-based evaluation metrics. Also, CBD introduces an independence assumption between chunks.

Sliding Window-based Decoding (SWBD) SWBD avoids such inter-chunk independence by sequentially translating each audio segment ($\mathbf{x}^n$), together with its corresponding previous source context ($\mathcal{C}^n_{\mathbf{x}}$). We distinguish two variants of SWBD. The first variant, SWBD, translates the concatenated segments and regards the last generated sentence as the translation of the current segment while discarding all other generations (Figure 2b). Note that this might introduce inconsistencies between the output produced at a time step, and the one used as target context in future time steps. By contrast, the second variant, SWBD-Cons, leverages the previously generated (up to C) sentences as a decoding constraint, based on which the model only needs to generate one sentence (Figure 2c).

In-Model Ensemble Decoding (IMED) We observe that SWBD still suffers from *misaligned translation*, where the translation of the current segment might contain information from previous segments. We introduce IMED to alleviate this issue as shown in Figure 2d. IMED extends SWBD-Cons by interpolating the document-level prediction (p^d) with the sentence-level prediction (p^s) as follows:

$$\lambda p^s_\theta\left(\mathbf{y}^n_t|\mathbf{y}^n_{<t},\mathbf{x}^n\right)+(1-\lambda)p^d_\theta\left(\mathbf{y}^n_t|\mathcal{C}\right), \quad (2)$$

where $\mathcal{C}=\{\mathcal{C}^n_{\mathbf{x}},\mathcal{C}^n_{\mathbf{y}},\mathbf{x}^n,\mathbf{y}^n_{<t}\}$, λ is a hyperparameter, $\mathbf{y}^n_t$ denotes the t-th target word in sentence $\mathbf{y}^n$, and both predictions are based on *the same model* θ. Intuitively, the sentence-level translation acts as a regularizer, avoiding the over- or under-translation. Note IMED with $\lambda=0$ corresponds to SWBD-Cons.

5 Experiments

5.1 Setup

We use the MuST-C dataset (Di Gangi et al., 2019) for experiments, which was collected from English TED talks and covers translations from English to 8 different languages, including German (De), Spanish (Es), French (Fr), Italian (It), Dutch (Nl), Portuguese (Pt), Romanian (Ro) and Russian (Ru). MuST-C offers a standard training, development and test set split for each language pair, with each dataset consisting of English audio, English transcriptions and their translations. Each training set contains transcribed speeches of ∼452 hours with ∼252K utterances on average. We report results on tst-COMMON, whose size ranges from 2502 (Es) to 2641 (De) utterances. We perform our major study on MuST-C En-De.

To construct acoustic features, for each audio segment, we extract 40-channel log-Mel filterbanks using overlapping windows of 25 ms and step size of 10 ms. We enrich these features with their first and second-order derivatives, followed by mean subtraction and variance normalization. Following Zhang et al. (2020a), we perform non-overlapping feature stacking to combine the features of three consecutive frames. All the texts are tokenized and truecased (Koehn et al., 2007), with out-of-vocabulary words handled by BPE segmentation (Sennrich et al., 2016), using 16K merging operations.

Model Settings and Evaluation Our context-aware ST follows Transformer base (Vaswani et al., 2017): 6 layers, 8 attention heads, and hidden/feed-forward size 512/2048. We use Adam ($\beta_1=0.9,\beta_2=0.98$) (Kingma and Ba, 2015) for parameter updates with label smoothing of 0.1. We use the same learning rate schedule as Vaswani et al. (2017) and set the warmup step to 4K. We apply dropout to attention weights and residual connections with a rate of 0.2 and 0.5, respectively. By default, we set $C=2$ and $\lambda=0.5$. Following (Zhang et al., 2020a), we apply AFS($\epsilon=-0.1,\beta=2/3$) to both temporal and feature dimensions for feature selection, which prunes out ∼84% speech encodings. We initialize our context-aware ST with the sentence-level Baseline, i.e. ST+AFS, and then finetune the model for 20K steps based on the concatenation method with a batch size of around 40K subwords.[4] We adopt beam search for decoding, with a beam size of 4 and length penalty of 0.6. We average the last 5 checkpoints for evaluation.

We measure general translation quality with tokenized case-sensitive BLEU (Papineni et al., 2002) and also report the detokenized one via *sacreBLEU* (Post, 2018)[5] for cross-paper comparison. We calculate BLEU based on sentences unless oth-

[4]Our experiments show that such initialization eases the learning of long inputs and improves the convergence of context-aware ST.

[5]signature: BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.3.6

ID	Model	BLEU	APT
1	Baseline (ST+AFS)	22.38 (27.40)	60.77
2	Ours + CBD	22.72 (27.95)	62.31
3	Ours + SWBD	22.70 (28.02)	**62.83**
4	Ours + SWBD-Cons	22.11 (27.98)	60.94
5	Ours + IMED	**22.86** (**28.03**)	62.56
6	1 + 20K-step finetuning	22.02 (27.00)	61.58
7	5 + $\lambda = 1.0$	22.42 (27.62)	61.96
8	1 + *lp* = 1.0	22.71 (27.77)	61.89
9	3 + *lp* = 1.0	**22.97** (**28.29**)	**63.51**
10	5 + *lp* = 1.0	22.94 (28.11)	62.76
11	3 w/o $\mathcal{C}_{\mathbf{y}}^{n}$	21.12 (26.17)	59.51
12	5 w/o $\mathcal{C}_{\mathbf{y}}^{n}$	20.72 (25.43)	58.18
13	3 w/o Baseline Initial.	21.75 (27.15)	62.29
14	5 w/o Baseline Initial.	21.97 (27.20)	62.08

Table 1: Case-sensitive tokenized BLEU and APT for different models and settings on MuST-C En-De test set. Numbers in bracket denote *document*-based BLEU. *lp*: the length penalty for beam search decoding. "w/o $\mathcal{C}_{\mathbf{y}}^{n}$": models that are trained without target-side context. Best results are highlighted in bold. Note $C = 2$, $\lambda = 0.5$ and *lp* = 0.6 by default.

erwise specified. We use APT (Miculicich Werlen and Popescu-Belis, 2017), the accuracy of pronoun translation, as an approximate proxy for document-level evaluation. Word alignment required by APT is automatically extracted via *fast align* (Dyer et al., 2013) with the strategy "*grow-diag-final-and*".

5.2 Results on MuST-C En-De

Does context improve translation? Yes, but the decoding method matters for context-aware ST. Table 1 summarizes the results. Our model with IMED outperforms Baseline by +0.48 BLEU (significant at $p < 0.05$)[6] and +1.79 APT (1→5), clearly showing the benefits from contextual modeling. Although SWBD-Cons yields worse sentence-based BLEU (-0.27, 1→4), it still beats Baseline in document-based BLEU (+0.58) and pronoun translation (+0.17 APT). The reason behind this inferior BLEU partially lies in misaligned translation (see Table 8 in Appendix for example). We observe that SWBD-Cons sometimes segments its output in a way that is misaligned to the reference segmentation. This also hurts CBD, where CBD produces mismatched sentences for around 1.8% cases. This is only a problem if we rely on the sentence-level alignment for BLEU, but not when we measure document-based BLEU (in brackets), where translations in one document are concatenated into a sequence for BLEU calculation. Overall, SWBD and IMED are more stable and perform the best, and SWBD surpasses Baseline by 2.06 APT (1→3). We will proceed with using IMED and SWBD for more reliable results with APT and later analysis.

Since we finetune our model based on the pre-trained Baseline, directly comparing with Baseline might be unfair. To offset its influence, we continue to train Baseline for the same 20K steps, following the settings in Section 5.1. Results show that this extra training (1→6) slightly deteriorates BLEU (-0.36) and only explains part of the improvement in APT (+0.81). Therefore, the gain brought by SWBD and IMED does not come from longer training. However, we do observe that initializing from the sentence-level Baseline benefits context-aware ST, compared to directly training context-aware ST from the AFS model (13→3, 14→4).

Apart from faster convergence and higher quality, another benefit of this finetuning is that the trained context-aware ST still carries the ability to translate individual sentences. Table 1 shows that using context-aware ST for sentence-level translation (1→7) yields similar BLEU to Baseline (+0.04) but surprisingly much better pronoun translation (+1.19), although it still underperforms SWBD and IMED. The fact that we can perform sentence-level ST using the same context-aware ST model indicates that it can be useful for ensembling, as confirmed by the effectiveness of IMED.

Upon closer inspection, we find that context-aware ST prefers to produce longer translations than Baseline. To control for the effects of output length on BLEU differences, we experiment with larger length penalty (*lp*: 0.6→1.0) to beam search. Results in Table 1 show that biasing the decoding greatly improves sentence-level ST (1→8), achieving performance on par with context-aware ST (when *lp* is 0.6) in terms of BLEU with similar translation lengths but still falling short of pronoun translation (-0.94 APT, 8→3). In addition, we observe that context-aware ST also benefits from decoding with larger length penalty, beating all sentence-level ST models (3→9, 5→10). Particularly, SWBD with *lp* of 1.0 delivers the best BLEU of 22.97 and APT of 63.51 (3→9). Note we adopt *lp* of 0.6 for the following experiments.

Does target-side context matter for context-aware ST? Yes, it matters a lot. By default, we utilize both source- and target-side context for contextual modeling. Removing the target-side part (also at training), as shown in Table 1 (11, 12), sub-

[6]We perform significance test using *bootstrap-hypothesis-difference-significance.pl* in *moses* (Koehn et al., 2007).

Model	BLEU	APT
SWBD	22.70	62.83
SWBD + Random $\mathcal{C}_{\mathbf{x}}^{n}$	22.31	61.16
IMED	22.86	62.56
IMED + Random $\mathcal{C}_{\mathbf{x}}^{n}$	21.83	59.95
IMED + Random $\mathcal{C}_{\mathbf{y}}^{n}$	21.99	60.01
IMED + Random $\mathcal{C}_{\mathbf{y}}^{n}$ & $\mathcal{C}_{\mathbf{y}}^{n}$	21.76	59.67

Table 2: Case-sensitive tokenized BLEU and APT for context-aware ST with random source/target context on MuST-C En-De test set. We report average performance over three runs with different random seeds. $C = 2, \lambda = 0.5$. Incorrect context hurts our model.

stantially weakens translation quality, even leading to worse performance than Baseline. Apart from offering direct target-side translation clues, we argue that the target-side context also enforces the context-aware ST to utilize the source-side context for translation, thus benefiting its training. This observation echoes with several previous studies on textual translation (Bawden et al., 2018; Huo et al., 2020; Lopes et al., 2020).

Does the model learn to utilize context? Yes. We answer this question by studying the impact of incorrect context on our model. We replace the correct source context with some random audio segments from the same document, and randomly select the target context from previous translations during decoding. Intuitively, the performance of our model should be intact if it ignores the context. Note that we trained our model with correct contexts but test it with random contexts here.

Results in Table 2 show that the randomized context, either source- or target-side, hurts the performance of our model in both BLEU and APT, similar to the findings in (Voita et al., 2018), and the translation of pronouns suffers more ($>$ -1.6 APT). Compared to SWBD, the incorrect context has more negative impact on IMED, resulting in worse performance than Baseline (Table 1), although IMED also uses sentence-level translation. We ascribe this to the target prefix constraint in IMED which makes translation errors at early decoding much easier to propagate. We observe that the incorrect target context acts similarly to its source counterpart under IMED, albeit its selection scope is much smaller (only limited to the translated segments), and combining both contexts leads to a slight but consistent performance degradation. These results demonstrate that our model indeed learns to use contextual information for translation.

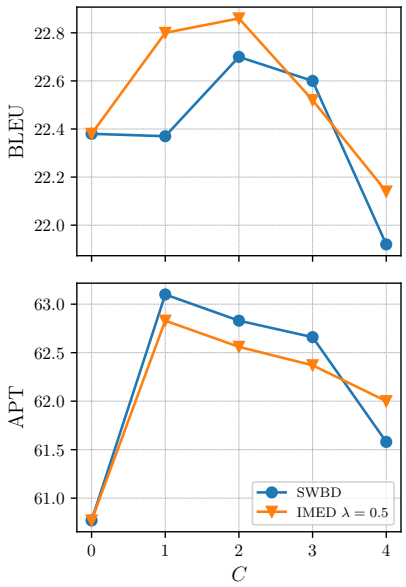

Figure 3: Case-sensitive tokenized BLEU (top) and APT (bottom) as a function of context size C on MuST-C En-De test set.

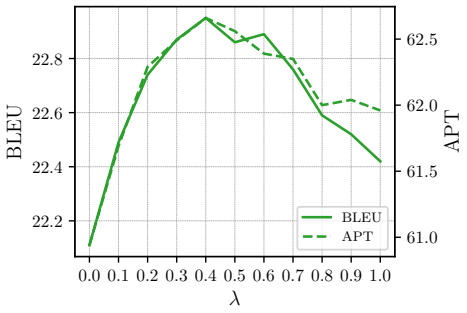

Figure 4: Case-sensitive tokenized BLEU (left y-axis) and APT (right y-axis) on MuST-C En-De test set when varying λ for IMED. Solid and dashed curves are for BLEU and APT, respectively. $C = 2$.

How much context sentences should we use? Although adding extra context provides more information, it makes learning harder: neural models often struggle with long sequences. Figure 3 shows the impact of context size on translation. We find that our models do not benefit from context size beyond 2 previous segments. Figure 3 also shows that the overall trend of the impact of C on BLEU and APT is similar for different decoding methods. Increasing C to 1 delivers the best APT, while context-aware ST achieves its best BLEU at $C = 2$. We use $C = 2$ for the following experiments.

Impact of λ on IMED. IMED heavily relies on the hyperparameter λ (Eq. 2) to control its preference between sentence-level and document-level decoding. Figure 4 shows its impact on translation

Model	ACC_{hp}
Baseline (ST+AFS)	48.93
Ours + SWBD	**49.90**
Ours + IMED	49.66
Ours + IMED $\lambda = 1.0$	48.77

Table 3: Translation accuracy of homophones (ACC_{hp}) on MuST-C En-De test set. $C = 2, \lambda = 0.5$.

Model	Random	Gold
Baseline (ST+AFS)	20.40	27.40
Ours + SWBD	21.83	28.02
Ours + IMED	**22.03**	**28.03**

Table 4: *Document*-level case-sensitive tokenized BLEU for different models on MuST-C En-De test set with erroneous audio segmentation. We report average BLEU over three runs; each run uses a different random seed to simulate segmentation errors. $C = 2, \lambda = 0.5$. *Random/Gold*: document-based BLEU when the random/gold segments are used.

quality, which clearly reveals a trade-off. The performance of IMED (BLEU and APT) reaches its peak at $\lambda = 0.4$, and decreases when λ becomes either smaller or larger. The optimal value of λ for IMED might vary greatly across different language pairs. It also shows some difference across evaluation sets (see Figure 7 in Appendix). In the following experiments, we will apply equal weighting ($\lambda = 0.5$), a common choice for model ensembles and not substantially worse than the optimum on this dataset.

Impact of context on homophone translation. Homophones (words that sound the same but hold different meanings, such as "*I*" vs. "*eye*" and "*would*" vs. "*wood*") and other acoustically similar words increase the learning difficulty of ST models compared to textual MT. To allow for automatic quantitative evaluation, we extract words from the MuST-C test set transcriptions which share the same phonemes with *Montreal Forced Aligner* (McAuliffe et al., 2017). We collect all homophones and evaluate their translation accuracy (ACC_{hp}) in the same way as APT.

Table 3 shows that context-aware ST outperforms Baseline by > 0.73 ACC_{hp}, where SWBD performs slightly better than IMED. After removing the document-level decoding, IMED ($\lambda = 1.0$) performance drops greatly, even underperforming Baseline. While we see some improvements to homophone translations, they are in the same relative range as general improvements from context. Anecdotal examples from manual inspection (see Table 7 in Appendix) indicate that context may at times help disambiguate acoustically similar forms, but that (near-)homophones still remain a salient source of translation errors.

Context improves the robustness of ST models to audio segmentation errors. In MuST-C, the audio is already well-segmented, with each segment corresponding to a short transcript. Nevertheless, natural audio, streaming speeches in particular, has no such segment boundaries, and how to partition audio itself is an active research area (Rangarajan Sridhar et al., 2013; Zhang and Zhang, 2020). Since ST models are often trained with gold segments, they inevitably suffer from segmentation errors at inference when the gold ones are unavailable.

The bottleneck mainly comes from the incompleteness of each segment, which, we argue, contextual information could alleviate. We simulate segmentation errors by randomly re-segmenting the audio in MuST-C En-De test set based on the given segment number. Especially, given an audio with N gold segments, we randomly re-segment it into N disjoint pieces, where each piece usually has different boundaries against its gold counterpart.[7] We evaluate different ST models with document-based BLEU.

Table 4 summarizes the results. Segmentation noise deteriorates translation quality for all ST models to a large degree ($>$ -6 BLEU). Compared to sentence-level ST, context-aware ST is less sensitive to those errors. In particular, our model with IMED yields a document-based BLEU of 22.03, substantially outperforming Baseline (by 1.63 BLEU). Our results also confirm the findings of Gaido et al. (2020).

Context benefits simultaneous translation. Simultaneous translation requires that we start decoding before receiving the whole audio input to minimize latency; operating on such short units increases ambiguity, and the model may be forced to predict future input to account for word order differences, which we hypothesize is easier with access to super-sentential context. We focus on segment-

[7] Note we intentionally keep the same segment number, N, in the simulated noisy segmentation, because this offers us a fair setup to analyze the impact of segmentation errors on the final translation when compared to the gold segmentation. This avoids the potential influence resulting from mismatched segment number. We leave the study of the model's robustness to genuine segmentation noises to future work.

Metric	Model	De	Es	Fr	It	Nl	Pt	Ro	Ru
BLEU↑	Baseline (ST+AFS)	22.38	27.04	33.43	23.35	25.05	26.55	21.87	14.92
	Ours + SWBD	22.70	27.12	34.23	23.46	25.84	26.63	23.70	15.53
	Ours + IMED	**22.86**	**27.50**	**34.28**	**23.53**	**26.12**	**27.37**	**24.48**	**15.95**
SacreBLEU↑	Baseline (ST+AFS)	22.4	26.9	31.6	23.0	24.9	26.3	21.0	14.7
	Ours + SWBD	22.7	27.0	32.4	23.0	25.7	26.4	22.8	15.4
	Ours + IMED	**22.9**	**27.3**	**32.5**	**23.1**	**26.0**	**27.1**	**23.6**	**15.8**
APT↑	Baseline (ST+AFS)	60.77	32.87	63.67	34.74	61.00	34.79	38.28	40.61
	Ours + SWBD	**62.83**	33.01	64.58	**35.20**	61.69	35.56	40.30	41.74
	Ours + IMED	62.56	**33.60**	**64.66**	**35.20**	**61.75**	**36.50**	**40.92**	**42.32**
ACC_{hp}↑	Baseline (ST+AFS)	48.93	43.85	56.96	**41.08**	50.73	43.64	47.07	30.80
	Ours + SWBD	**49.90**	43.73	57.30	40.04	51.48	44.03	47.66	**32.67**
	Ours + IMED	49.66	**44.66**	**57.76**	40.62	**52.07**	**45.42**	**48.49**	32.56

Table 5: Results on MuST-C for 8 language pairs. We set $C = 2, \lambda = 0.5$. Numbers in bold are the best results.

Model	BLEU↑	DAL↓	NE↓
Baseline (ST+AFS)	21.02	3.97	1.72
Ours + SWBD	21.86	3.82	1.95
Ours + SWBD-Cons	21.98	**3.75**	**1.59**
Ours + IMED	**22.55**	3.91	1.64

Table 6: Simultaneous translation results (BLEU, DAL and NE) for different models on MuST-C En-De test set. $C = 2, \lambda = 0.5$.

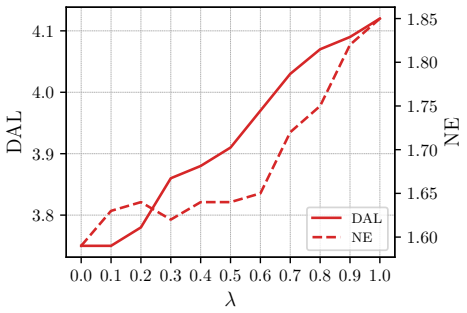

Figure 5: DAL (left y-axis) and NE (right y-axis) as a function of λ for IMED on MuST-C En-De test set in simultaneous translation setting. Solid and dashed curves are for DAL and NE, respectively. $C = 2$. $\lambda \to 0.0$: document-level decoding; $\lambda \to 1.0$: sentence-level decoding.

level E2E simultaneous translation, and adopt the re-translation method (Niehues et al., 2016; Arivazhagan et al., 2020b,a) where we translate the source input segment from scratch after every 1 second. For training, we finetune each model for extra 20K steps with a 1:1 mix of full-segment and prefix pairs, following Arivazhagan et al. (2020a). We construct the prefix pairs by uniformly selecting an audio prefix length and then proportionally deciding the target prefix length based on the sentence length. Note that the context inputs in our model are still full segments/sentences. We adopt tokenized BLEU, differentiable average lagging (DAL), and normalized erasure (NE) to evaluate the translation quality, latency and stability, respectively, following Arivazhagan et al. (2020a). Note DAL and NE are measured based on words.

Results in Table 6 show that context-aware ST improves translation quality (> +0.84 BLEU) and reduces translation latency (> -0.06 DAL) regardless of the decoding method. It also enhances translation stability when the target prefix constraint is applied (> -0.08 NE, SWBD-Cons & IMED). SWBD performs worse in NE, because it allows changes in the translation of context which increases instability. Overall, context provides extra information to the translation model, before the E2E ST models see the whole input, which benefits simultaneous translation.

Figure 5 further illustrates how context impacts simultaneous translation. With the increase of sentence-level decoding ($\lambda \to 1.0$), IMED produces higher DAL and NE, i.e. worse quality. We ascribe the reduction of latency and stability in our model to the inclusion of contextual information.

5.3 Results on Other Language Pairs

Table 5 summarizes the results for all 8 translation pairs covered by MuST-C. Overall, our model obtains improvements over most metrics and language pairs, despite their different language characteristics. Out of 8 languages, our model performs relatively worse on Es and It with smaller BLEU gains and even negative results in ACC_{hp}. By contrast, our model yields the largest improvement on Ro. In particular, our model with IMED achieves a detokenized BLEU of 23.6 on En-Ro, surpassing the state-of-the-art result 22.2 (Zhao et al., 2020) reported so far.

6 Conclusion and Future Work

Our experiments confirm the effectiveness of context-aware modeling for end-to-end speech translation. With concatenation-based contextual modeling and appropriate decoding method, we observe positive impact of context on translation. Context-aware ST improves general translation quality in BLEU, and also helps pronoun and homophone translation. ST models become less sensitive to (artificial) audio segmentation errors with context. In addition, context also improves simultaneous translation by reducing latency and erasure. We observe overall positive results over different languages and evaluation metrics on the MuST-C corpus.

In the future, we will investigate more dedicated neural architectures to handle long-form speech input. While we relied on a dataset with sentence segmentation in this work, we are interested in removing the reliance on segmentation at inference time to implement the full-fledged streaming translation scenario.

Acknowledgements

We thank the reviewers for their insightful comments. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements 825460 (ELITR). Rico Sennrich acknowledges support of the Swiss National Science Foundation (MUTAMUR; no. 176727).

References

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020a. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Isabelle Te, Wolfgang Macherey, Pallavi Baljekar, and George Foster. 2020b. Re-translation strategies for long form, simultaneous, spoken language translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7919–7923. IEEE.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Xinyi Cai and Deyi Xiong. 2020. A test suite for evaluating discourse phenomena in document-level neural machine translation. In *Proceedings of the Second International Workshop of Discourse Processing*, pages 13–17, Suzhou, China. Association for Computational Linguistics.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019. Adapting Transformer to End-to-End Spoken Language Translation. In *Proc. Interspeech 2019*, pages 1133–1137.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2020. Sdst: Successive decoding for speech-to-text translation. *arXiv preprint arXiv:2009.09737*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2020. Contextualized Translation of Automatically Segmented Speech. In *Proc. Interspeech 2020*, pages 1471–1475.

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until

the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1342–1352, Doha, Qatar. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Liane Kirsten Guillou. 2016. *Incorporating pronoun function into statistical machine translation*. Ph.D. thesis, University of Edinburgh.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island, Korea. Association for Computational Linguistics.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Adrià Giménez Pastor, Joan Albert Silvestre-Cerdà, Pau Baquero-Arnal, Jorge Civera Saiz, and Alfons Juan. 2020. Direct segmentation models for streaming speech translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2599–2611, Online. Association for Computational Linguistics.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020a. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Xutai Ma, Yongqiang Wang, Mohammad Javad Dousti, Philipp Koehn, and Juan Pino. 2020b. Streaming simultaneous speech translation with augmented memory transformer. *arXiv preprint arXiv:2011.00033*.

Evgeny Matusov, Dustin Hillard, Mathew Magimai-Doss, Dilek Hakkani-Tür, Mari Ostendorf, and Hermann Ney. 2007. Improving speech translation with automatic boundary prediction. In *Eighth Annual Conference of the International Speech Communication Association*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference*

on Empirical Methods in Natural Language Processing, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2016. Dynamic transcription for low-latency speech translation. In *Interspeech 2016*, pages 2513–2517.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 230–238, Atlanta, Georgia. Association for Computational Linguistics.

Annette Rios, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738, Online. Association for Computational Linguistics.

Deyi Xiong and Min Zhang. 2013. A topic-based coherence model for statistical machine translation. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, page 977–983. AAAI Press.

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2020a. Adaptive feature selection for end-to-end speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2533–2544, Online. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2020b. On sparsifying encoder outputs in sequence-to-sequence models. *arXiv preprint arXiv:2004.11854*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020c. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online. Association for Computational Linguistics.

Ruiqing Zhang and Chuanqiang Zhang. 2020. Dynamic sentence boundary detection for simultaneous translation. In *Proceedings of the First Workshop*

on Automatic Simultaneous Translation, pages 1–9, Seattle, Washington. Association for Computational Linguistics.

Chengqi Zhao, Mingxuan Wang, and Lei Li. 2020. Neurst: Neural speech translation toolkit. *arXiv preprint arXiv:2012.10018*.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3983–3989. International Joint Conferences on Artificial Intelligence Organization. Main track.

A Impact of C and λ on Dev Set

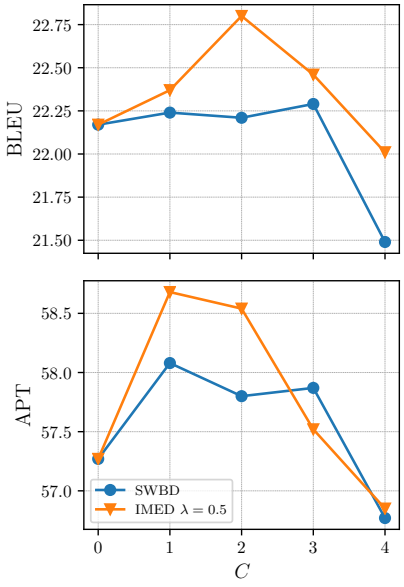

Figure 6: Case-sensitive tokenized BLEU (top) and APT (bottom) as a function of context size C on MuST-C En-De **dev set**.

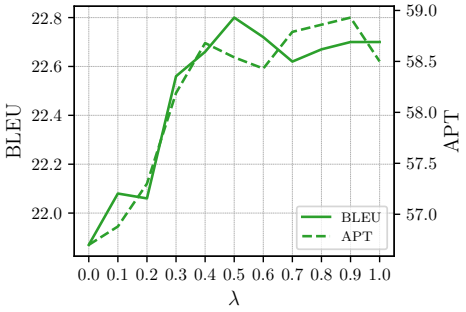

Figure 7: Case-sensitive tokenized BLEU (left y-axis) and APT (right y-axis) on MuST-C En-De **dev set** when varying λ for IMED. Solid and dashed curves are for BLEU and APT, respectively. $C = 2$.

Results in Figure 6 and 7 show that the optimal value of C and λ also differs across evaluation sets. Overall, setting $C = 2$ and $\lambda = 0.5$ offers us decent performance. Note again, we selected these configurations for generality and simplicity rather than its being optimal.

B Case Study on Homophone Translation

C Examples for Misaligned Translation

Context	I remember my first fire.
Source	I was the second volunteer on the **scene**, so there was a pretty good chance I was going to get in.
Reference	Ich war der zweite Freiwillige an der **Brandstelle**, ich hatte also recht gute Chancen hinein zu können.
Baseline	Ich war der zweite Freiwillige auf der **CNU**, also war ich ziemlich gut darin.
Ours + SWBD	Ich war der zweite Freiwillige auf der **CNN**, also gab es eine ziemlich gute Chance, dass ich sie bekommen würde.
Ours + SWBD-Cons	Ich war der zweite Freiwillige auf dem **CNN**, also gab es eine ziemlich gute Chance, dass ich sie bekommen würde.
Ours + IMED	Ich war der zweite Freiwillige auf dem **CNN**, also war ich ziemlich gut darin, dass ich ihn kriegen würde.
Context	The Human Genome Project started in 1990, and it took 13 years.
Source	It **cost** 2.7 billion dollars.
Reference	Es **kostete** 2,7 Milliarden Dollar.
Baseline	Es **kostet** 2,7 Milliarden Dollar. *(EN: costs)*
Ours + SWBD	Es **kostete** 2,7 Milliarden Dollar.
Ours + SWBD-Cons	Es **kostete** 2,7 Milliarden Dollar.
Ours + IMED	Es **kostet** 2,7 Milliarden Dollar. *(EN: costs)*

Table 7: Examples of translation errors due to confusion with near-homophones (bold) from the MuST-C En-De test set.

(1)	Source	She asked the monk, "Why is it that her hand is so warm and the rest of her is so cold?" "Because you have been holding it since this morning," **he said. "You have not let it go."**
	Reference	Sie fragte den Mönch: "Wieso ist ihre Hand so warm und der Rest von ihr ist so kalt?" "Weil Sie sie seit heute morgen halten", **sagte er. "Sie haben sie nicht losgelassen."**
	Translation	Sie fragte den Monat: "Warum ist ihre Hand so warm?" Und der Rest von ihr ist so kalt, weil ihr seit diesem Morgen das hält.
(2)	Source	If there is a sinew in our family, it runs through the women.
	Reference	Wenn es in unserer Familie ein Band gibt, dann verläuft es durch die Frauen.
	Translation	**Er sagte: "Sie haben es nicht geschafft, loszulassen."**

Table 8: Example of misaligned translation for SWBD-Cons from the MuST-C En-De test set. The translation for the second segment (2) actually aligns with the first one (1), as highlighted in bold.